

Sensing Higgs boson cascade decays through memory

Christoph Englert^{1,*} Malcolm Fairbairn^{2,†} Michael Spannowsky^{3,‡} Panagiotis Stylianou^{1,§} and Sreedevi Varma^{2,||}

¹*SUPA, School of Physics & Astronomy, University of Glasgow, Glasgow G12 8QQ, United Kingdom*

²*Theoretical Particle Physics and Cosmology, Kings College London,
London WC2R 2LS, United Kingdom*

³*Institute for Particle Physics Phenomenology, Durham University, Durham DH1 3LE, United Kingdom*



(Received 10 September 2020; accepted 24 October 2020; published 30 November 2020)

Beyond the Standard Model scenarios with extensions of the Higgs sector typically predict new resonances that can undergo a series of cascade decays to detectable Standard Model particles. On one hand, sensitivity to such signatures will contribute to the full reconstruction of the extended Higgs potential if a new physics discovery will be made. On the other hand, such cascade decays could be dominant decay channels, thus being potentially the best motivated signatures to achieve a new physics discovery in the first place. In this work, we show how the long short-term memory that is encoded in the cascade decays' phenomenology can be exploited in discriminating the signal from the background, where no such information is present. In parallel, we demonstrate for theoretically motivated scenarios that such an approach provides improved sensitivity compared to more standard analyses, where only information about the signal's final state kinematics is included.

DOI: [10.1103/PhysRevD.102.095027](https://doi.org/10.1103/PhysRevD.102.095027)

I. INTRODUCTION

The search for new physics beyond the Standard Model (SM) of particle physics is the main driver of the phenomenology programme at the Large Hadron Collider (LHC). The current negative outcome of beyond the SM (BSM) searches seems to suggest that new degrees of freedom are either too heavy or too weakly coupled to be experimentally accessible at this stage in the LHC programme.

If new physics is related to the top quark and Higgs boson sector, as is expected in most concrete ultraviolet (UV) completions of the SM that tackle the shortcomings of the SM such as insufficient CP violation or TeV scale naturalness, another phenomenologically interesting avenue arises: new exotic scalar bosons could be dominantly produced through SM-Higgs like gluon fusion, Fig. 1(a). If this production mode is relevant as a consequence of sizable Yukawa couplings (or phases), unitarity typically implies a large decay probability into top quarks when kinematically

accessible.¹ However, it is known [3–8] that large accidental interference of QCD-induced $t\bar{t}$ production with the scalar state can create a significant distortion of the on-shell resonance signal. When including constraints from dark matter searches, low energy experiments, flavor physics, 125 GeV Higgs signal strength measurements and exotic Higgs searches as done in Ref. [9], motivated UV completions such as the two-Higgs-doublet model (2HDM) (for a review, see [10]) are forced into parameter regions that are particularly impacted by these interference effects.

This could mean that new physics is already present at the energy scales presently being explored by the LHC, yet interference renders the signal difficult to detect in the best motivated $t\bar{t}$ channel. If this is the case, sensitivity to these models can be restored using di-Higgs final states. While these final states can be enhanced by constructive signal-signal interference in concrete UV extensions of the Higgs sector [9], the significantly reduced sensitivity to such signatures will mean that new physics discoveries will be pushed into the LHC's high luminosity (HL) phase.

In scenarios with a richer scalar phenomenology, multi-Higgs production from cascade decays of a new scalar degree of freedom into a 125 GeV SM Higgs h and another BSM scalar boson are possible. These signatures arise in, e.g., the next-to-minimal 2HDM [11] (N2HDM) with sizable cross sections and provide an important

*christoph.englert@glasgow.ac.uk

†malcolm.fairbairn@kcl.ac.uk

‡michael.spannowsky@durham.ac.uk

§p.stylianou.1@research.gla.ac.uk

||sreedevi.varma@kcl.ac.uk

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

¹Decays into massive quarks are typically further enhanced due to symmetry considerations such as custodial isospin [1] or CP properties of the new scalar state [2].

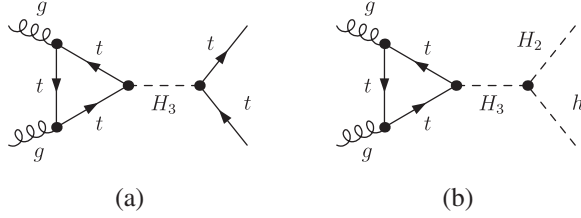


FIG. 1. Representative gluon fusion diagrams for the production of an exotic scalar H_i and subsequent decay into either $H_3 \rightarrow t\bar{t}$ (a) or scalar decays $H_3 \rightarrow H_2 h$ (b).

phenomenological input for the reconstruction of the extended Higgs potential.² In scenarios like the complex 2HDM, such signatures directly probe alignment of the 125 GeV Higgs boson with fluctuations around the electroweak vacuum-independent from decoupling of additional states [13] and are therefore theoretically well motivated. Depending on the mass of the final state exotic Higgs boson, such cascade signatures also arise in the next-to-minimal supersymmetric standard model (NMSSM) [14,15].

In this work, we focus on decays of heavy scalars $H_3 \rightarrow H_2 h$, Fig. 1(b), where we identify the 125 GeV SM Higgs boson as the lightest scalar degree of freedom, $H_1 = h$. We are specifically interested in the parameter region where $m_{H_3} > m_{H_2} + m_h$ and $m_{H_2} > 2m_t$, i.e., the region of parameter space where the decay

$$H_3 \rightarrow H_2 h, \quad \text{with} \quad H_2 \rightarrow t\bar{t}, h \rightarrow b\bar{b} \quad (1)$$

is open and sizable. Such final states are experimentally challenging due to b-jet combinatorics and a significant amount of missing energy that renders the reconstruction of resonances difficult. While distinct kinematical correlations that are induced by the cascade decay structure can be accessed through observables like M_{T2} of Refs. [16,17], an analysis strategy based on rectangular combinations of collider observables might be too restrictive to obtain the highest statistical signal yield for a given background rejection.³ In parallel, the particular hierarchy of the branchings of Eq. (1) induces a “time scale” for the signal events which is not present for the contributing background. While phenomenological analyses aim to perform an appropriate clustering of the final state’s kinematics on a statistical level, they typically do so without accessing the event’s memory imprint directly.

Fingerprinting the relevance of this memory for signal vs background discrimination is the focus of this work. We will access this memory by means of recurrent neural networks (RNNs) and show its relevance by comparing

²See also [12] for a discussion of Higgs cascade decays in the context of the two-singlet extension of the SM.

³Shower and event deconstruction are alternative all-information approaches to discriminate hadronically decaying top quarks and Higgs bosons from backgrounds [18–20].

this setup against other signal-background discrimination methodologies.

This paper is organized as follows. In Sec. II A, we quickly motivate the use of RNNs for the physics problem that we study in this paper and outline our analysis setup in Sec. II B. Section II C gives an overview of the different strategies that we employ, and Sec. II D compares the efficacy of those strategies. We summarize and conclude in Sec. III.

II. CASCADING MEMORY

A. General remarks and context

RNNs are networks designed to train on a sequence of time ordered events rather than spatially distributed values. In the case under consideration in this work, we employ an RNN to identify the flow of particles in the showering and branching after a collision event.

The architecture of the recurrent network allows it to connect a piece of information to the previous piece of information learnt, the classic example being connecting the end of the sentence to the beginning.

RNNs are widely used in translation (many-to-many), music generation (one-to-many), and sentiment classification or reading joined up handwriting (many-to-one) applications. Depending on the task, recurrent neural networks might have different architectures. Our case is most similar to the structure of a many-to-one situation where the many represents the string of events as particles decay into each other and the one is the nature of the hard particles created in the initial event in the collider.

The RNN is a machine learning network where nodes are replaced by units, individual gates rather like logic gates but made up of algorithms consisting of fixed combinations of algebraic and smooth activation functions, as well as weights connecting those functions. These units are then distributed across the network in the same way nodes would form a normal neural network, with freedom in the choice of architecture, e.g., the number of layers and the number of units in each layer, etc.

The input is split into time ordered components like the consecutive words in a sentence and each word is fed in an ordered fashion into the first layer of units, consecutively from left to right. Within each layer of the network, each unit produces an activation and an output which are fed to the next layer; however, the output is also fed sideways to the right in the same layer so that the input into each unit contains information about the previous words in the sentence only.

In our case, the words are replaced by the parameters of jets/leptons/missing energy with the time ordering replaced by p_T ordering.

The weights are then varied during training using the same gradient descent algorithms used for normal neural network training, and the global minimum of the cost/error function is searched for (and hopefully obtained).

Updating weights requires propagation through the network of derivatives of both the cost/error function and the codependence of weights upon each other. The required chain-rule multiplication of many such derivatives increases the risk of gradients vanishing or blowing up. Hence, memorizing a longer sequence is a challenging task in traditional RNNs. Long short-term memory (LSTM) units can be used to construct a particular class of RNN, which address these issues and remembers information for a longer period [21]. LSTM and closely related gated recurrent unit (GRU) [22] networks have a gated structure that regulates the passage of information through the unit. The LSTM architecture therefore stabilizes the way that the units change their behavior as weights are updated.

Applications of RNNs to jet physics have emerged in recent years. LSTMs have been in use for flavor tagging [23,24], substructure studies [25–27], hardware analysis [28,29], and event-level classifiers [30]. Various deep learning techniques to classify light-flavored and heavy-flavored jets are compared in [23]. Particle tracks and vertices are used as the classifying features of the network. A comparison study comparing RNNs to deep neural networks (DNNs), LSTMs and outer recursive networks, while exploiting their prowess in tracing the full event history, was performed in this paper. The classification of jets (up quark initiated vs down quark initiated) using their electric charge was attempted in [26]. Convolutional, recurrent, and recursive neural networks were used to train the network. The approach followed in [30] is to investigate the analogy between the way RNNs perform natural language processing to training on jet physics. Jets derived from sequential clustering algorithms are fed into the network and used for classification purpose.

B. Event data and preprocessing

In this work, we consider on the cascade decay signature

$$pp \rightarrow H_3 \rightarrow (H_2 \rightarrow t\bar{t} \rightarrow \ell^+ \ell'^- b\bar{b} + \cancel{E}_T) + (h \rightarrow b\bar{b}), \quad (2)$$

i.e., we feed in the two leptons, four b-jets and the missing energy as the discriminating features of the signal. Pseudorapidity η , azimuthal angle ϕ , transverse momentum p_T , and energy E parameters are used to pass the information into the network. We focus on 13 TeV collisions. The signature of two leptons, four b-tagged jets, and missing energy arise when the tops decay to b quarks, leptons, and neutrinos, thus providing a range of correlations and a cluster history that is not (fully) present for the contributing background processes, which include

- (i) $pp \rightarrow t\bar{t}b\bar{b}$,
- (ii) $pp \rightarrow t\bar{t}(Z \rightarrow b\bar{b})$,
- (iii) $pp \rightarrow t\bar{t}(h \rightarrow b\bar{b})$,
- (iv) $pp \rightarrow bb\bar{b}\bar{b}W^+W^-$,

TABLE I. Inclusive cross sections for background processes at a 13 TeV LHC, where $t\bar{t}b\bar{b}$, $t\bar{t}h$, and $t\bar{t}Z$ normalizations include K-factors 1.8 [31], 1.17 [32], and 1.2 [33], respectively.

Process	Cross section (fb)
$pp \rightarrow t\bar{t}b\bar{b}$	1215.050
$pp \rightarrow t\bar{t}(h \rightarrow b\bar{b})$	22.007
$pp \rightarrow t\bar{t}(Z \rightarrow b\bar{b})$	6.096
$pp \rightarrow bb\bar{b}\bar{b}W^+W^-$	2.561
$pp \rightarrow bb\bar{b}\bar{b}ZZ$	0.014

(v) $pp \rightarrow bb\bar{b}\bar{b}ZZ$.

Out of these, the $t\bar{t}b\bar{b}$ production is by far the most dominant contribution; see Table I.

The signal is modeled with FeynRules [34,35] and we generate signal and background events with MadEvent [36–38] and MadSpin [39,40]. The generated events are showered with PYTHIA8 [41] and outputted in the HepMC format [42]. We use FastJet [43,44] for clustering jets, interfaced through the reconstruction mode of MadAnalysis [45–48]. All jets are clustered with the anti-kT algorithm [49] of radius 0.4 with the requirement that they have a transverse momentum of

$$p_T(j) > 20 \text{ GeV} \quad (3)$$

and a pseudorapidity of

$$|\eta(j)| < 4.5. \quad (4)$$

B-jets are selected with efficiency of $\epsilon = 0.8$ and in the central part of the detector within

$$|\eta(j_b)| < 2.5. \quad (5)$$

The final state leptons are selected if

$$p_T(\ell) > 5 \text{ GeV} \quad \text{and} \quad |\eta(\ell)| < 2.5. \quad (6)$$

Subsequently, we impose isolation criteria, where a lepton is considered isolated if the total p_T of the jets within the light lepton's cone radius $R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} = 0.3$ is less than 20% of the lepton's transverse momentum $p_T(\ell)$. The event is accepted if exactly two leptons and four b-tagged jets are identified, otherwise vetoed. The missing transverse momentum is evaluated as the opposite to the four-momenta sum of jet and lepton tracks in the plane perpendicular to the beam and its magnitude is considered as the missing transverse energy.

The production of $t\bar{t}b\bar{b}$ provides the largest background contribution. $bb\bar{b}\bar{b}Z$ is many orders of magnitude smaller, and we will not consider it further. Event numbers are rescaled to a sum of 69000 events before passed to the neural network with an equal number of signal events for training.

C. Architectures

Before we turn to an application of the RNN strategy to analyze an actual N2HDM scenario, we would like to quantify the information gain that becomes available by using RNN as opposed to other strategies that do not directly access the memory of the signal event decay chain. To this end, the data were trained using two different networks to assess the importance of long- and short-term memory on the classification score, i.e., we compare the performance of the RNN with that of a DNN. Models are built using TensorFlow2.1 [50] and trained using NVIDIA GeForce GTX1080TI and RTX 2080TI on the CUDA10.2 platform [51]. We use 81% of events for training, 9% for validation, and 10% for testing.

The RNN network is trained for many different architectures—we vary the number of GRU (LSTM) layers from one to nine, while the number of RNN units also vary from 10 to 100. Default parameters are used in the GRU/LSTM units, and Tanh activation is applied to the units while the sigmoid function is used as the recurrent activation. Weights are initialized using Glorot [52] uniform initializers in the GRU/LSTM units while orthogonal initializers are used in the recurrent states. A dropout parameter of 0.1 is used between the layers to avoid overfitting.

The DNN is trained using identical hyperparameters (without dropout). The number of layers and the number of units are varied as in the RNN. Weights of the DNN are also initialized using Glorot [52], and ReLU [53] activation was given to the layers.

Both the RNN and the DNN are optimized using the Adam [54] algorithm using categorical cross-entropy with a learning rate of 0.001 and default beta parameters, which are the exponential decay rates for the first- (0.9) and second-moment estimates (0.99), respectively. The networks are trained for 100 epochs with early patience of 10. The output layer is activated using softmax [55] activation to obtain the class probability (binary cross-entropy with the output layer activated using a sigmoid function produces similar results).

D. Performance comparison and physics

To check whether an LSTM/GRU network provides additional sensitivity in events with long decay chains, we perform a scan over signal configurations with masses M_{H_3} , $M_{H_2} \in [410, 950]$ GeV and H_3 width at 10% and 30% of M_{H_3} . LSTM and DNN networks are trained on each signal sample evaluating the background rejection of the network when minimum signal efficiencies of 10% and 30% are required. The model-independent scan over masses considers directly the MC-truth information without showering or realistic selection criteria to clarify the *a priori* usability of both types of neural networks for the considered cascade decays. The LSTM network shows an improved performance compared to the DNN, especially

when larger signal efficiencies are required as shown in Fig. 2. The kinematical observables of the final state particles largely depend on the mass of the different resonant structures in the BSM. Therefore, the networks are able to better discriminate from the SM background for larger H_2 , H_3 masses that lead to a more pronounced cascade decay phenomenology. This leads to the reduction of the background and opens up the possibility of excluding points of the parameter space of concrete scenarios. Including effects from showering and hadronization (which creates additional sources of missing energy from meson decays), and realistic acceptance criteria Eqs. (3)–(6), we show sensitivity projections for the HL-LHC (3/ab) as a function of $M_{H_{3,2}}$ in Fig. 3. As the background rapidly falls with M_{H_3} mass hypothesis, we are sensitive to smaller cross sections at higher mass.

Given the RNN network's enhanced sensitivity to the cascade decay's phenomenology, we can further discuss its relevance for motivated scenarios beyond the generic scan of Fig. 3. To this end, we focus on the N2HDM as a prototype scenario that predicts the signature of Eq. (2). Relevant coupling points are obtained by scanning the parameter space of the N2HDM using ScannerS [11,56–59] and requiring the branching ratios of the scalars $\text{BR}(H_3 \rightarrow H_2 h)$, $\text{BR}(H_2 \rightarrow t\bar{t})$, and the pseudoscalar $\text{BR}(A \rightarrow t\bar{t})$ to be larger than 0.5.⁴ To demonstrate the sensitivity that is available through the GRU/LSTM setup, we focus on a N2HDM parameter point with $M_{H_2} = 480$ GeV, $M_{H_3} = 722$ GeV, and widths 4.9 and 45 GeV for H_2 and H_3 , respectively. This point has a cross section of 3.43 fb and passes the branching requirements with $\text{BR}(H_3 \rightarrow H_2 h) = 0.52$. QCD corrections for the signal are included via reweighting to the Higgs Cross Section Working Group values [65] (see also [66,67]) for the scale choice of $\mu = M_{H_3}/2$. Again, we include the effects of showering as well as additional sources of missing energy that arise from hadronization and meson decays.

As usual in machine learning, the performance of the classifiers is visualized using the receiver operating characteristics (ROC) curves. Histograms of the class probability values are also plotted to check how well the predicted probability values are separated for a given classification threshold. Again, we train the networks with different hyperparameters and ROC curves of networks with good performance on both training and validation data, which are shown in Fig. 4. We find that the RNN always shows a slightly better discrimination. This leads us to the conclusion that the splitting history encoded in the event indeed provides relevant information that allows one to discriminate the background from the signal in a slightly more nuanced way. While the cascade decay leaves discriminating features in the final state kinematic

⁴CMS and ATLAS are searching for charged [60,61] and neutral Higgs bosons [62–64], which can provide additional sensitivity to this parameter region.

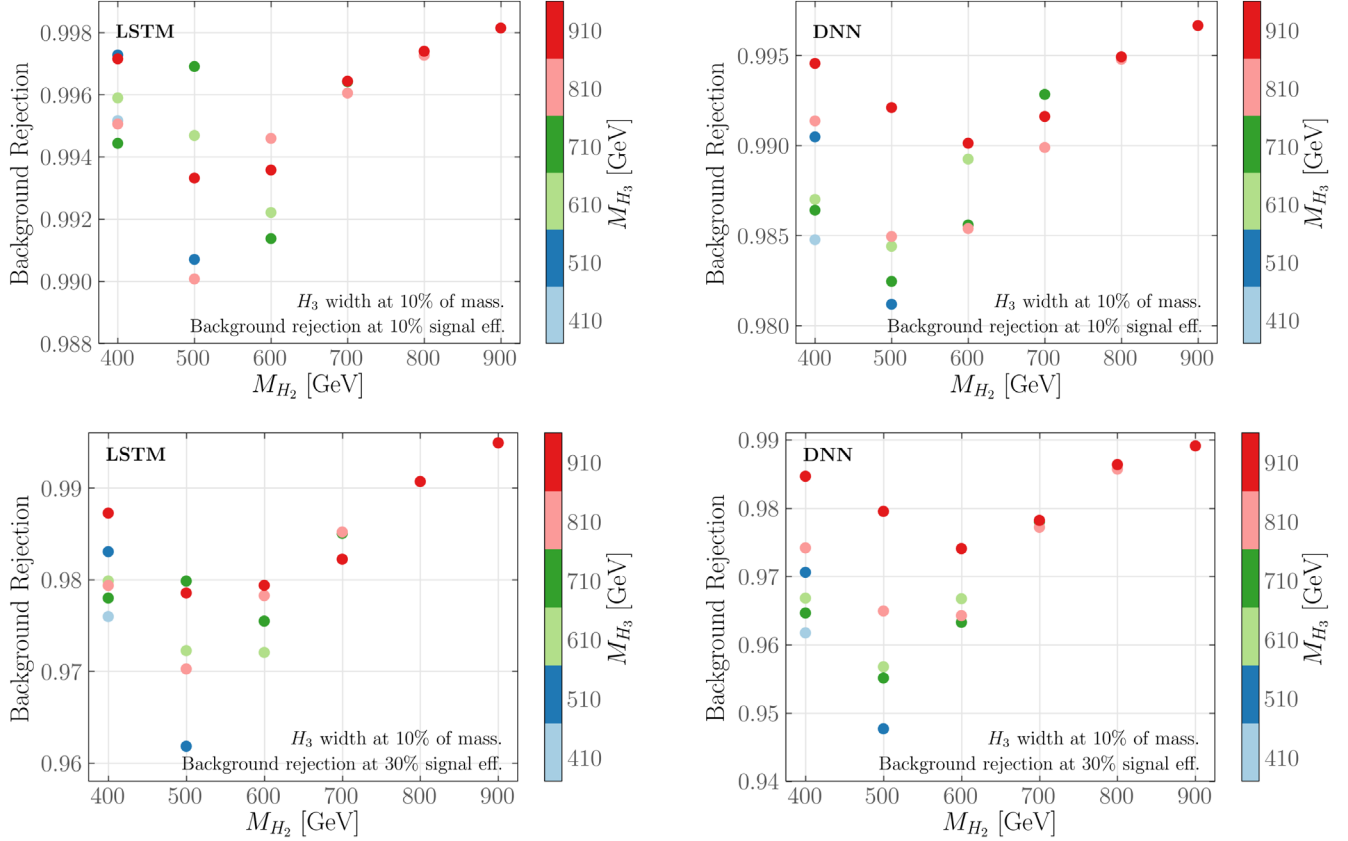


FIG. 2. Figures showing the background rejection of LSTM and DNN networks for signal efficiencies of 10% and 30%. The signal is generated with different masses for each case and $M_{H_3} > M_{H_2} + m_h$. The width of H_3 is set to 10% of its mass, while the width of H_2 is calculated assuming 100% branching into top quarks. Note that these assumptions predominantly influence the normalization and not the efficiency of the networks. For each signal case, the network is trained along with the background events for a luminosity of 500/fb. The LSTM network has one LSTM layer of 45 units, a dropout rate of 0.1, and learning rate of 0.001, while the DNN network has two fully connected layers of 80 units and a learning rate of 0.001. The mass scans are performed using Monte Carlo truth particles before showering and particle reconstruction. The networks use 10000 events split into training, testing, and validation sets. Runs with the width of H_3 set to 30% of its mass produced comparable results.

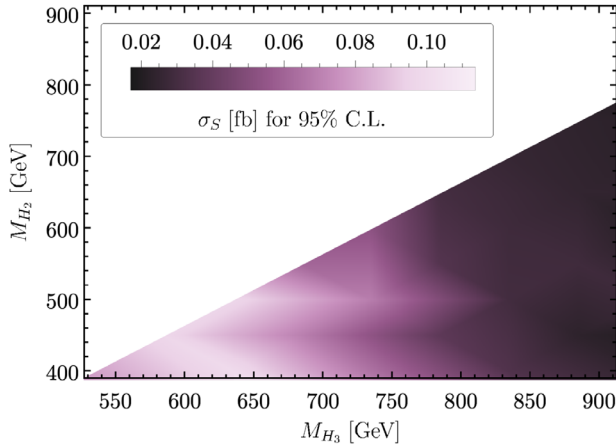


FIG. 3. Sensitivity in the $M_{H_3} - M_{H_2}$ plane displayed as the final signal cross section σ_S required to achieve $S/\sqrt{B} = 2$ (95% C.L.), using the LSTM framework described in the text. We scan over the mass parameters with the requirement $M_{H_3} > M_{H_2} + 125$ GeV and fixed branching ratios $\text{BR}(H_3 \rightarrow H_2 h) = 0.5$ and $\text{BR}(H_2 \rightarrow t\bar{t}) = 1$.

information which are used by the DNN to perform the classification, the quicker convergence of the RNN setup demonstrates that this information is more efficiently learned through an adapted architecture that reflects the branching hierarchy directly. The initial loss for the DNN in Fig. 4 is higher than that for the RNN, but this is a function of architecture—the initial loss typically becomes comparable to LSTM/GRU initial losses as the number of layers varies from 1 up to 10. It is interesting to note that the cascade decay structure is crucial to the improved performance of the RNN setup. For instance, we can consider the separation of $t\bar{t}Z$ from $t\bar{t}h$ production. For our NN input data, the discrimination is driven by the invariant $b\bar{b}$ mass, while both processes have a comparable resonance structure. In this case, a comparison of RNN and DNN architectures does not single out the RNN as a better adapted approach.

A strong test of our setup is its capability to isolate the resonance structures from the provided input data in the presence of the significant degradation from missing

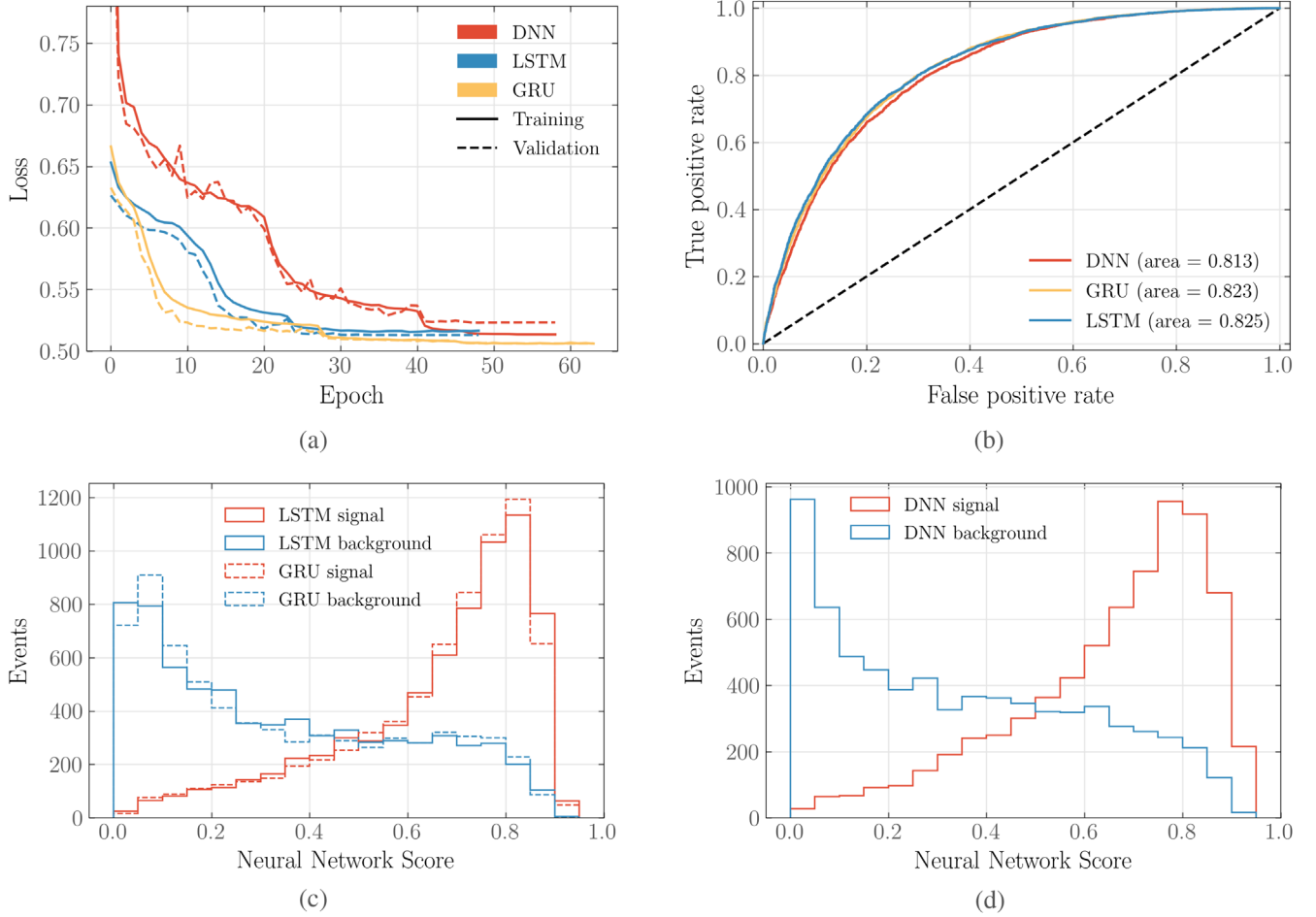


FIG. 4. (a) Loss curves for training and validation data for LSTM, GRU, and DNN along with (b) ROC curves for each case for the N2HDM benchmark masses (see text). Class probability values for the (c) RNN cases and for (d) DNN. The LSTM (GRU) consists of one layer of 45 units which result in comparable performance, while the DNN is built with two dense layers of 80 units. The latter provides slightly poorer discrimination of signal against background and requires more epochs to minimize the loss function.

energy. We define a reconstructed M_{H_3} by adding the four-momenta of the two leptons and the four b-tagged jets of highest p_T as well as the missing transverse momentum. Similarly, M_{H_2} is defined from two leptons and a pair of b-tagged jets incompatible with the 125 GeV Higgs mass, 125 ± 10 GeV. We can use these definitions to check that we indeed get a peaklike structures after training and selection. The reconstructed masses are shown in Fig. 5, before and after the application of an LSTM network to select events. Although the resonance structure becomes significantly distorted due the sizable missing energy that arises from a range of sources, the resonance peaks are visible, and the backgrounds are significantly reduced in a signal-like selection.

To determine the sensitivity quantitatively, we perform an analysis of signal and background rates after the application of the LSTM. The cross sections obtained after the LSTM selection which is chosen to maintain a large σ_S/σ_B (σ_S and σ_B are the signal and background cross sections after selection, respectively). This is done to minimize impact of background systematics which we

neglect in this study. Subsequently, we perform a pseudo-measurement by evaluating the signal (background) number of events S (B) at an extrapolated integrated luminosity of 3/ab and determining the significance S/\sqrt{B} . For an LSTM network of one layer with 45 units, we obtain a significance of 5.3 based on a rate of $S/B \simeq 0.09$. Performing the same analysis with a DNN network of two dense layers with 80 units each, the significance is $S/\sqrt{B} = 4.1$ at $S/B \simeq 0.08$. This shows that the DNN is slightly more vulnerable to background systematics, while the GRU/LSTM architecture is essential to claim a new physics discovery in this channel at the HL-LHC. Finally, for comparison, we additionally perform a simple cut-and-count analysis to conclude our comparison of different approaches. Besides the selection criteria, additional cuts are imposed on the missing energy requiring $\cancel{E}_T > 30$ GeV. The search region is further constrained by applying cuts on the transverse momentum of final state particles. The four b-jets must satisfy staggered cuts $p_T(b_1) > 100$ GeV, $p_T(b_2) > 70$ GeV, $p_T(b_3) > 65$ GeV,

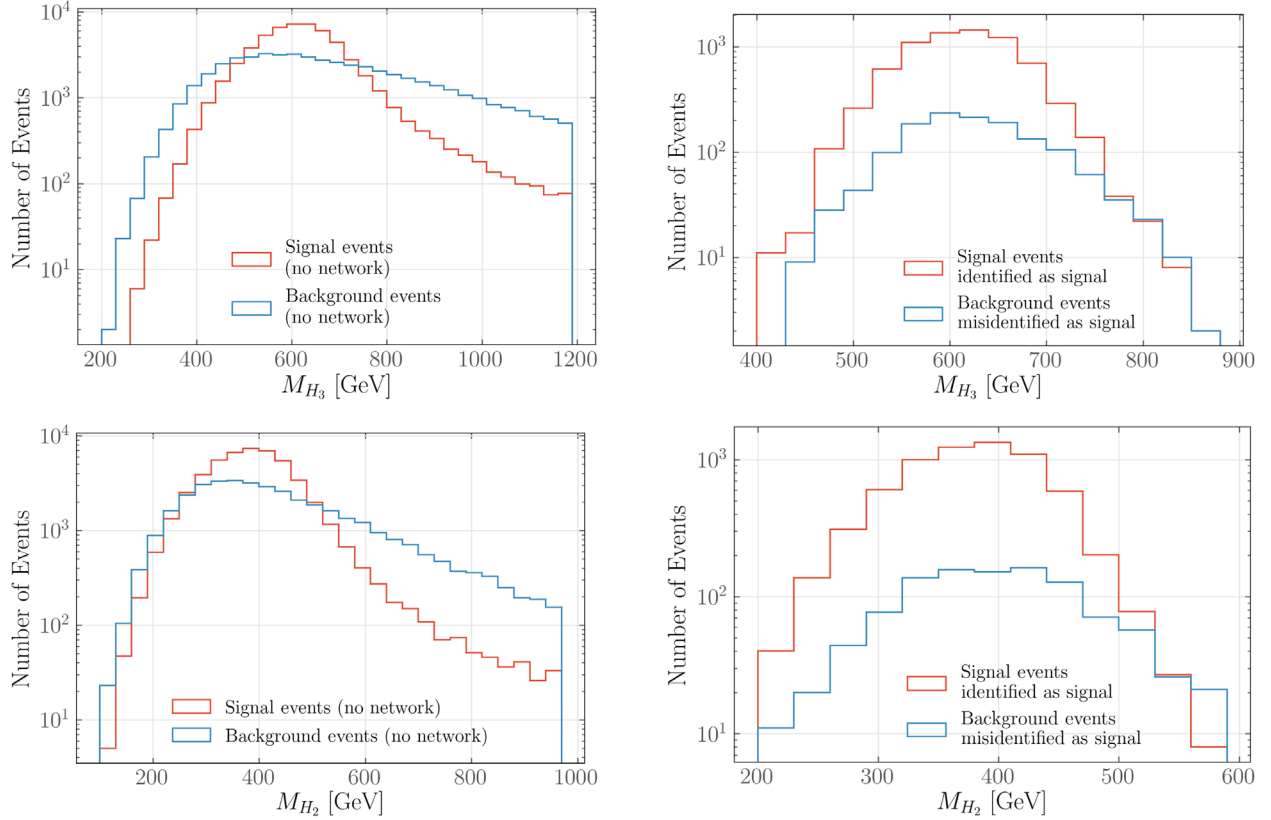


FIG. 5. Example histograms with and without the LSTM neural network for an N2HDM set of couplings with $M_{H_2} = 480$ GeV, $M_{H_3} = 722$ GeV, and widths 4.9 and 45 GeV for H_2 and H_3 , respectively. The LSTM network used had one LSTM layer of 45 units, a dropout rate of 0.1, and learning rate of 0.001.

and $p_T(b_4) > 50$ GeV, while for the leptons, similarly $p_T(\ell_1) > 30$ GeV and $p_T(\ell_2) > 10$ GeV were imposed. A Higgs compatible pair is reconstructed by requiring the invariant mass of a pair of b-jets to be within 125 ± 10 GeV. If more than one possible pair is identified, the one with the smallest separation ΔR is selected and if no candidate pair is found the event is vetoed. The reconstructed Higgs must satisfy $p_T(h) > 120$ GeV and the invariant mass of the remaining two b-jets is restricted to $m_{bb} > 80$ GeV. The aforementioned cuts result in a smaller S/B ratio compared to the network approaches, evaluated as 0.04 which corresponds to a significance of 2.1. This poorer performance highlights the relevance of using as much information as possible in discriminating signal from background as given by the LSTM/GRU and DNN networks to gain sensitivity to new physics scenarios such as the N2HDM at the LHC.

III. DISCUSSION AND CONCLUSIONS

The search for new physics beyond the Standard Model in scenarios with exotic scalars in cascade decays can be subject to interference effects in the best motivated top final states if additional scalars are heavy. In this case, multi-Higgs production can come to the rescue, as it will provide

additional sensitivity to the “standard” SM-like Higgs searches. Furthermore, gaining sensitivity to such decays is crucial for the reconstruction of the underlying microscopic theory. Particularly motivated in this context are decays of a heavy Higgs state into a pair of different mass Higgs bosons, one of which is the 125 GeV state. Such signatures probe particular aspects of the models’ UV structure such as 2HDM alignment or an extension of the 2HDM scalar sector, thus also helping to discriminate between different model hypotheses if a discovery is made.

In this work, we have exploited the memory imprinted by cascade decay patterns of a heavy state through a chain of decay steps into SM matter. We have demonstrated that RNNs which access this memory in a particularly adapted way exhibit superior discriminative power than “ordinary” DNNs, which would need to learn the decay steps indirectly through correctly pairing final state objects. In general, this results in a slightly reduced sensitivity of DNN networks for the considered physics case. In parallel, the DNN performance results from a longer learning period while the RNNs pick up the available information rapidly. To highlight the physical relevance of this approach, we have considered a parameter point of the N2HDM model space that could be observed in the cascade decay channel using the RNN approach with a significance of over 5σ at the HL-LHC.

The RNN architecture is particularly relevant for this parameter point to be able to claim a discovery at the LHC. While we have focused on a particular decay chain, our results can be expected to generalize to the other UV scenarios such as the NMSSM where the scalar mass scales are different, and H_2 would lie below the 125 GeV boson with direct decays $H_2 \rightarrow b\bar{b}$. We leave this for future work.

ACKNOWLEDGMENTS

The work of C.E. was supported by the UK Science and Technology Facilities Council (STFC) under Grant

No. ST/P000746/1 and by the IPPP Associateship Scheme. The work of M. S. was supported by the STFC under Grant No. ST/P001246/1. The work of P. S. was supported by an STFC studentship under Grant No. ST/T506102/1. The work of M.F. was funded partly by the STFC Grant No. ST/L000326/1. The work of S. V. and M. F. were also supported by the European Research Council under the European Union's Horizon 2020 programme (ERC Grant No. 648680 DARK-HORIZONS). S. V. was the recipient of a Sir Richard Trainor Scholarship at the start of her Ph.D.

-
- [1] J. F. Gunion, R. Vega, and J. Wudka, *Phys. Rev. D* **42**, 1673 (1990).
 - [2] A. Djouadi, *Phys. Rep.* **459**, 1 (2008).
 - [3] K. J. F. Gaemers and F. Hoogeveen, *Phys. Lett.* **146B**, 347 (1984).
 - [4] D. Dicus, A. Stange, and S. Willenbrock, *Phys. Lett. B* **333**, 126 (1994).
 - [5] S. Jung, J. Song, and Y. W. Yoon, *Phys. Rev. D* **92**, 055009 (2015).
 - [6] W. Bernreuther, P. Galler, C. Mellein, Z. G. Si, and P. Uwer, *Phys. Rev. D* **93**, 034032 (2016).
 - [7] M. Carena and Z. Liu, *J. High Energy Phys.* **11** (2016) 159.
 - [8] A. Djouadi, J. Ellis, A. Popov, and J. Quevillon, *J. High Energy Phys.* **03** (2019) 119.
 - [9] P. Basler, S. Dawson, C. Englert, and M. Mhlleitner, *Phys. Rev. D* **101**, 015019 (2020).
 - [10] G. C. Branco, P. M. Ferreira, L. Lavoura, M. N. Rebelo, M. Sher, and J. P. Silva, *Phys. Rep.* **516**, 1 (2012).
 - [11] M. Mhlleitner, M. O. P. Sampaio, R. Santos, and J. Wittbrodt, *J. High Energy Phys.* **03** (2017) 094.
 - [12] T. Robens, T. Stefaniak, and J. Wittbrodt, *Eur. Phys. J. C* **80**, 151 (2020).
 - [13] B. Grzadkowski, H. E. Haber, O. M. Ogreid, and P. Osland, *J. High Energy Phys.* **12** (2018) 056.
 - [14] P. Basler, S. Dawson, C. Englert, and M. Mhlleitner, *Phys. Rev. D* **99**, 055048 (2019).
 - [15] S. Baum and N. R. Shah, *arXiv:1904.10810*.
 - [16] C. G. Lester and D. J. Summers, *Phys. Lett. B* **463**, 99 (1999).
 - [17] A. Barr, C. Lester, and P. Stephens, *J. Phys. G* **29**, 2343 (2003).
 - [18] D. E. Soper and M. Spannowsky, *Phys. Rev. D* **84**, 074002 (2011).
 - [19] D. E. Soper and M. Spannowsky, *Phys. Rev. D* **87**, 054012 (2013).
 - [20] D. E. Soper and M. Spannowsky, *Phys. Rev. D* **89**, 094005 (2014).
 - [21] S. Hochreiter and J. Schmidhuber, *Neural Comput.* **9**, 17351780 (1997).
 - [22] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Doha, Qatar, 2014), pp. 1724–1734, <https://www.aclweb.org/anthology/D14-1179>.
 - [23] D. Guest, J. Collado, P. Baldi, S.-C. Hsu, G. Urban, and D. Whiteson, *Phys. Rev. D* **94**, 112002 (2016).
 - [24] M. Aaboud *et al.* (ATLAS Collaboration), Report No. ATL-PHYS-PUB-2017-003, 2017.
 - [25] S. Egan, W. Fedorko, A. Lister, J. Pearkes, and C. Gay, *arXiv:1711.09059*.
 - [26] K. Fraser and M. D. Schwartz, *J. High Energy Phys.* **10** (2018) 093.
 - [27] A. Butter *et al.*, *SciPost Phys.* **7**, 014 (2019).
 - [28] M. Wielgosz, A. Skoczeń, and M. Mertik, *Nucl. Instrum. Methods Phys. Res., Sect. A* **867**, 40 (2017).
 - [29] M. Wielgosz, A. Skoczeń, and M. Mertik, *arXiv:1702.00833*.
 - [30] G. Louppe, K. Cho, C. Becot, and K. Cranmer, *J. High Energy Phys.* **01** (2019) 057.
 - [31] A. Bredenstein, A. Denner, S. Dittmaier, and S. Pozzorini, *Phys. Rev. Lett.* **103**, 012002 (2009).
 - [32] D. de Florian *et al.* (LHC Higgs Cross Section Working Group), *arXiv:1610.07922*.
 - [33] ATLAS and CMS Collaborations, *CERN Yellow Rep. Monogr.* **7**, 140 (2019).
 - [34] N. D. Christensen and C. Duhr, *Comput. Phys. Commun.* **180**, 1614 (2009).
 - [35] A. Alloul, N. D. Christensen, C. Degrande, C. Duhr, and B. Fuks, *Comput. Phys. Commun.* **185**, 2250 (2014).
 - [36] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer, *J. High Energy Phys.* **06** (2011) 128.
 - [37] P. de Aquino, W. Link, F. Maltoni, O. Mattelaer, and T. Stelzer, *Comput. Phys. Commun.* **183**, 2254 (2012).
 - [38] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, *J. High Energy Phys.* **07** (2014) 079.
 - [39] S. Frixione, E. Laenen, P. Motylinski, and B. R. Webber, *J. High Energy Phys.* **04** (2007) 081.
 - [40] P. Artoisenet, R. Frederix, O. Mattelaer, and R. Rietkerk, *J. High Energy Phys.* **03** (2013) 015.

- [41] T. Sjstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, *Comput. Phys. Commun.* **191**, 159 (2015).
- [42] M. Dobbs and J. B. Hansen, *Comput. Phys. Commun.* **134**, 41 (2001).
- [43] M. Cacciari, G. P. Salam, and G. Soyez, *Eur. Phys. J. C* **72**, 1896 (2012).
- [44] M. Cacciari and G. P. Salam, *Phys. Lett. B* **641**, 57 (2006).
- [45] E. Conte, B. Fuks, and G. Serret, *Comput. Phys. Commun.* **184**, 222 (2013).
- [46] E. Conte, B. Dumont, B. Fuks, and C. Wymant, *Eur. Phys. J. C* **74**, 3103 (2014).
- [47] B. Dumont, B. Fuks, S. Kraml, S. Bein, G. Chalons, E. Conte, S. Kulkarni, D. Sengupta, and C. Wymant, *Eur. Phys. J. C* **75**, 56 (2015).
- [48] E. Conte and B. Fuks, *Int. J. Mod. Phys. A* **33**, 1830027 (2018).
- [49] M. Cacciari, G. P. Salam, and G. Soyez, *J. High Energy Phys.* **04** (2008) 063.
- [50] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, [arXiv:1603.04467](https://arxiv.org/abs/1603.04467).
- [51] J. Nickolls, I. Buck, M. Garland, and K. Skadron, *Queueing Syst. Theory Appl.* **6**, 40 (2008).
- [52] X. Glorot and Y. Bengio, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, edited by Y. W. Teh and M. Titterton (PMLR, Chia Laguna Resort, Sardinia, Italy, 2010); Vol. 9 of *Proceedings of Machine Learning Research*, pp. 249–256, <http://proceedings.mlr.press/v9/glorot10a.html>.
- [53] V. Nair and G. E. Hinton, in *Proceedings of the 27th International Conference on International Conference on Machine Learning* (Omnipress, USA, 2010), ICML'10, pp. 807–814, ISBN: 978-1-60558-907-7, <http://dl.acm.org/citation.cfm?id=3104322.3104425>.
- [54] D. P. Kingma and J. Ba, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [55] J. S. Bridle, in *Neurocomputing*, edited by F. F. Soulié and J. Hérault (Springer Berlin Heidelberg, Berlin, Heidelberg, 1990), pp. 227–236, ISBN 978-3-642-76153-9.
- [56] R. Coimbra, M. O. P. Sampaio, and R. Santos, *Eur. Phys. J. C* **73**, 2428 (2013).
- [57] P. M. Ferreira, R. Guedes, M. O. P. Sampaio, and R. Santos, *J. High Energy Phys.* **12** (2014) 067.
- [58] R. Costa, M. Mhlleitner, M. O. P. Sampaio, and R. Santos, *J. High Energy Phys.* **06** (2016) 034.
- [59] M. Mhlleitner, M. O. P. Sampaio, R. Santos, and J. Wittbrodt, [arXiv:2007.02985](https://arxiv.org/abs/2007.02985).
- [60] M. Aaboud *et al.* (ATLAS Collaboration), *J. High Energy Phys.* **11** (2018) 085.
- [61] A. M. Sirunyan *et al.* (CMS Collaboration) *J. High Energy Phys.* **07** (2020) 126.
- [62] A. M. Sirunyan *et al.* (CMS Collaboration), *J. High Energy Phys.* **03** (2020) 055.
- [63] A. M. Sirunyan *et al.* (CMS Collaboration), Report No. CMS-PAS-HIG-17-027.
- [64] M. Aaboud *et al.* (ATLAS Collaboration), *Phys. Rev. Lett.* **119**, 191803 (2017).
- [65] S. Dittmaier *et al.* (LHC Higgs Cross Section Working Group), [arXiv:1101.0593](https://arxiv.org/abs/1101.0593).
- [66] S. Dawson, *Nucl. Phys.* **B359**, 283 (1991).
- [67] D. Graudenz, M. Spira, and P. M. Zerwas, *Phys. Rev. Lett.* **70**, 1372 (1993).